NICE National Institute for Health and Care Excellence

Evidence generation plan for artificial intelligence (AI) technologies to aid opportunistic detection of vertebral fragility fractures: early value assessment

Implementation support

Published: 14 October 2025

www.nice.org.uk

Contents

1 Purpose of this document

NICE's early value assessment of artificial intelligence (AI) technologies to aid opportunistic detection of vertebral fragility fractures recommends that BriefCase-Triage, CINA-VCF Quantix, HealthVCF, HealthOST and IB Lab FLAMINGO can be used in the NHS during the evidence generation period. Stakeholders should consider whether the technologies are likely to remain available on the UK market and supported by their companies before generating evidence to address the evidence gaps. Evidence generation should preferably be on technologies that will still be available in the NHS after the evidence generation period.

This plan outlines the evidence gaps and what real-world data needs to be collected for a NICE review of the technologies again in the future. It is not a study protocol but suggests an approach to collecting the information needed to address the evidence gaps. Evidence generated through other study approaches will also be considered. For assessing comparative treatment effects, well-conducted randomised controlled trials are the preferred source of evidence.

The company is responsible for ensuring that data collection and analysis takes place.

Guidance on commissioning and procurement of the technology will be provided by NHS England.

NICE will withdraw all or part of the guidance if a company does not meet the conditions in section 4 on monitoring.

At the end of the evidence generation period (3 years), the companies should submit the evidence to NICE in a format that can be used for decision making. NICE will review all the evidence and assess whether the technologies can be routinely adopted in the NHS.

2 Evidence gaps

This section describes the evidence gaps, why they need to be addressed and their relative importance for future committee decision making.

The committee will not be able to make a positive recommendation without the essential evidence gaps (see section 2.1) being addressed. The companies can strengthen the evidence base by addressing as many other evidence gaps (see section 2.2) as possible. This will help the committee to make a recommendation by ensuring it has a better understanding of the patient or healthcare system impact of the technologies.

2.1 Essential evidence for future committee decision making

Impact of the artificial intelligence technologies on health-related quality of life

The committee noted that there was limited evidence about how the artificial intelligence (AI) technologies affect health-related quality of life in the short term of at least 12 months. EQ-5D-3L is the preferred tool for measuring this outcome. At committee, a review of the evidence from the technologies showed that the utility gain had the largest impact on the cost-effectiveness results for these technologies.

Resource use

More information is needed on how using the technologies would affect resource use during and after implementation, to help the committee understand their long-term resource use impacts. Key areas that will help to address this evidence gap are:

- long-term resource use costs, such as number and extent of treatments and number of hospital appointment or visits
- the downstream impacts of using the technologies on the NHS, such as:
  - the number of people referred for spine X-ray or dual-energy X-ray absorptiometry

(DEXA) scan, or

  - the number of people receiving medication for osteoporosis

- the time taken to process diagnostic images by reporting practitioners, including additional reviews by specialists.

Ideally, information about implementation, technology acquisition and maintenance costs and payment models could also be collected.

Impact of using AI technologies on the NHS care pathway

A key part of the committee discussion was around the impact of AI technologies on the NHS care pathway for fragility fractures and osteoporosis. For example, changes in the fracture liaison services diagnosis and treatment routes may be needed to accommodate the AI technology. Collecting evidence on this will help the committee understand how using the technologies will affect care in the NHS.

Failure rates and diagnostic accuracy of the AI technologies ideally compared with NHS standard care

The committee noted that the failure rates and diagnostic accuracy outcomes for NHS standard care were not reported adequately. More evidence is needed on the failure rates and diagnostic accuracy of the AI technologies compared with current NHS care.

2.2 Evidence that further supports committee decision making

Diagnostic accuracy of the AI technologies in people under 50 years

The failure rate and diagnostic accuracy outcomes for people younger than 50 years and at risk of VFF, for example people with long-term corticosteroid use or malignancy in the vertebrae, were not reported adequately. More evidence is needed on the failure rates and diagnostic accuracy of the AI technologies when used in these groups.

3 Approach to research

3.1 Evidence gaps and ongoing studies

Table 1 summarises the evidence gaps and ongoing studies that might address them. Information about evidence status is derived from the external assessment report. More information on the studies in the table can be found in the supporting documents.

Table 1 Evidence gaps and ongoing studies

Evidence gap	BriefCase-Triage	CINA-VCF Quantix	HealthOST	IB Lab FLAMINGO
Health-related quality-of-life impacts of AI technologies	No evidence	No evidence	No evidence	No evidence
Resource use	Limited evidence	Limited evidence Ongoing study	Limited evidence	Limited evidence Ongoing study
Impact of using AI technologies on NHS care pathway	No evidence	No evidence	No evidence	No evidence
Failure rates and diagnostic accuracy of AI technologies ideally compared with NHS standard care	Limited evidence	Limited evidence	Limited evidence	Limited evidence
Healthcare professional experience and acceptability of AI technologies	No evidence	No evidence	No evidence	No evidence

Abbreviation: AI, artificial intelligence.

3.2 Data sources

NICE's real-world evidence framework provides detailed guidance on assessing the suitability of a real-world data source to answer a specific research question.

The Fracture Liaison Service Database (FLS-DB) could potentially support this research.

This database contains patient-level data on secondary fracture prevention in England and Wales, collected as part of the Falls and Fragility Fracture Audit Programme. It includes much of the data needed to address the evidence gaps, such as individual patient outcome data items, identification of fragility fracture, and the bone therapy recommended.

The Diagnostic Imaging Dataset (DID) is a national collection of detailed information about diagnostic imaging tests done in the NHS. It could be used to address some evidence gaps, specifically around failure rates and diagnostic accuracy of the AI technologies ideally compared with NHS standard care. The data for DID is extracted from local Radiology Information Systems and submitted monthly.

Patient-level data from FLS-DB and DID can be linked to other datasets, such as NHS Digital's Hospital Episode Statistics. This could support the evaluation of longer-term outcomes such as adverse events and resource use in the NHS, such as further hospital appointments and referral for treatment.

The quality and coverage of real-world data collections are of key importance when used in generating evidence. Active monitoring and follow up through a central coordinating point is an effective and viable approach for ensuring good-quality data with broad coverage.

3.3 Evidence collection plan

Most of the evidence gaps can be addressed through a real-world before-and-after implementation study. A retrospective service-evaluation study using available databases is also proposed to evaluate the failure rates and diagnostic accuracy of the AI technologies ideally compared with NHS standard care evidence gap.

Real-world before-and-after implementation study

This type of study can assess an intervention's impact by comparing measurements from before and after its implementation. In this instance the impact of the AI technologies on health-related quality of life, resource use and the care pathway would be assessed in both phases. Once the technologies have been implemented, data about their failure rates and diagnostic accuracy in a real-world setting can also be collected.

After an enrolment period, data collection should be long enough for sufficient follow-up.

The AI technologies should then be implemented and data collected to assess their impact, after leaving a period of time to account for learning effects.

While this study could be done at a single centre, it should ideally be implemented across NHS trusts with and without fracture liaison services and replicated across multiple centres. This could show how the AI technology can be implemented across a range of settings, representative of the variety in the NHS. In addition, this would allow for adjusting for site variation when analysing outcome data. Outcomes may reflect other changes that occur over time in the population, unrelated to the interventions. Additional robustness can be achieved by:

- collecting data in a centre that has not implemented an AI technology but is as similar as possible (in terms of clinical practice and patient characteristics) to one that has, or
- ideally through a stepped-wedge design.

This could help control for changes in diagnosis and treatment rates over time that might have occurred anyway.

Developers could initially do a 'silent evaluation' (see Kwong et al. 2022) before full deployment into services. This approach allows the technology to be used in a real-world setting without any influence on clinical decision making until it is fully deployed. This approach can be used to:

- understand whether the technology can be deployed safely (including in subpopulations)
- understand how it might have influenced decision making (for example, onward referrals)
- collect some relevant data items (for example, failure rate or number of indeterminate findings).

In order to mitigate the committee's concerns about the resource impact of implementing the technologies (see section 3.24 of the guidance), initial uses should be on a small scale. Wider rollout may be possible within the period of evidence generation if, and when, it becomes clear that the resource impact of the technologies is manageable.

Retrospective study

Current NHS standard care failure rates and diagnostic accuracy could be evaluated using data from FLS-DB and DID in a retrospective study. These failure rates and diagnostic accuracy findings should then be compared with those of implemented AI technologies.

For AI technologies indicated for use in all adults over 18 years, data from FLS-DB and DID should be used to evaluate failure rates and diagnostic accuracy in people younger than 50 years and at risk of VFF, for example people with long-term corticosteroid use or malignancy in the vertebrae.

3.4 Data to be collected

The following information has been identified for collection:

Retrospective study

- Patient demographics, including age, sex and ethnicity
- Diagnostic accuracy
- Accuracy when used by different reporting practitioners
- Failure rate or rate of inconclusive AI reports
- Number of missed vertebral fragility fractures
- Rate of missed fracture-related further injury
- Proportion of people who need further imaging
- Conditions that may complicate imaging (for example, obesity or scoliosis)
- Health-related quality-of-life data, ideally collected with the EQ-5D-3L questionnaire
- Details of the technology (software name, version and configuration settings)
- Image details (including anatomical location, projection when considering X-rays and manufacturer of CT or X-ray machine).

Real-world before-and-after implementation study

- Patient demographics, including age, sex and ethnicity
- Conditions that may complicate imaging (for example, obesity or scoliosis)
- Health-related quality-of-life data, ideally collected with the EQ-5D-3L questionnaire
- Detail of the technology (software name, version and configuration settings)
- Image details (including anatomical location, projection when considering X-rays and manufacturer of CT or X-ray machine)
- Time taken to process and report image with AI assistance
- Additional reviews by specialists
- Subsequent scanning, for example spinal X-ray and dual-energy X-ray absorptiometry (DEXA) scans
- Time to diagnosis
- Time to further referral or treatment
- Pre- and post-diagnosis treatment status
- Number of hospital appointments, including referrals to fracture clinics and orthopaedic assessments
- Impact of complications, such as number of hospital admissions and emergency department visits.

Data collection should follow a predefined protocol, and a quality assurance process should be put in place to ensure the integrity and consistency of data collection. See NICE's real-world evidence framework, which provides guidance on the planning, conduct and reporting of real-world evidence studies.

Information about the technology

Information about how the technology was developed should also be reported, including:

- the characteristics of the patient data used in the AI training datasets

- the version tested
- how the effect of future updates will be monitored.

The AI training datasets should include younger people, ethnic minorities and people with comorbidities or who have had previous treatment. This will ensure that the technologies can be analysed, tested or validated in diverse patient populations. See the NICE evidence standards framework for digital health technologies.

3.5 Evidence generation period

The evidence generation period should be 3 years. This will be enough time to set up and implement the AI technologies, collect the necessary data and analyse it.

3.6 Following best practice in study methodology

Following best practice when conducting studies is paramount to ensuring the reliability and validity of the research findings. Adhering to rigorous guidelines and established standards is crucial for generating credible evidence that can ultimately improve patient care. NICE's real-world evidence framework details some key considerations.

For an early value assessment, a key factor to consider as part of the informed consent process is ensuring that patients (and their carers, as appropriate) understand that data will be collected to address the evidence gaps identified in section 2. Where applicable, this should take account of NICE's guidance on shared decision making.

4 Monitoring

NICE will contact the companies:

- within 6 months of publication of this plan to confirm agreements are in place to generate the evidence
- annually to confirm that the data is being collected and analysed as planned.

The companies should tell NICE as soon as possible of anything that may affect ongoing evidence generation, including:

- any substantial risk that the evidence will not be collected as planned
- new safety concerns
- the technology significantly changing in a way that affects the evidence generation process.

If a company's data collection is expected to end later than planned, that company should contact NICE to arrange an extension to the evidence generation period. NICE reserves the right to withdraw all or part of the guidance if data collection is delayed, or if it is unlikely to resolve the evidence gaps.

5 Minimum evidence standards

The evidence considered by NICE primarily focuses on diagnostic accuracy, failure rates, training and implementation of the artificial intelligence (AI) technologies in varied non-NHS settings. All the technologies that have been recommended for use in the NHS during the evidence generation period have some implementation experience in the NHS. The companies did not report any safety concerns. The technologies demonstrated potential to be cost effective for opportunistic detection of vertebral fragility fractures. But, more information is still needed to fully understand the benefits they may provide. More evidence is needed on:

- the AI technologies' impacts on health-related quality of life
- the diagnostic accuracy and failure rates of the AI technologies compared with NHS standard care
- longer-term clinical and resource-use impacts of AI technologies to the NHS
- the impact on the care pathway.

This evidence is essential to future NICE decision making. It will also potentially inform the optimum use and implementation of AI technologies for vertebral fragility fracture detection in the NHS.

6 Implementation considerations

The following considerations around implementing the evidence generation process have been identified through working with system partners:

Evidence generation

- The companies should collect and analyse outcome data carefully to ensure that important subgroups are included in the studies, such as people:
  - under 50
  - with obesity where the field of view for a diagnostic image may include more surrounding tissue
  - with cancer
  - with osteogenesis imperfecta.

Equalities

- During implementation of artificial intelligence (AI) technologies, having limited access to fracture liaison services, fragility fractures and osteoporosis treatments in the NHS may drive health inequalities. This could worsen regional inequalities, particularly for people living in deprived areas.

System and implementation considerations

- It is unknown how these AI technologies might impact the skills of the healthcare professional in detecting vertebral fragility fractures. Care should be taken to ensure that healthcare professionals are not deskilled by over-reliance on AI technologies.
- The companies should ideally provide or support training for healthcare professionals in using the technologies.
- The current regulatory approval for HealthVCF is due to expire in 2028, so the technology is unlikely to be available on the UK market after 2028. The preference is

for evidence to be generated using HealthOST while it is used in the NHS, because this is the technology that will be more widely available in the future. Data from HealthVCF may not be generalisable to HealthOST (see section 3.13 of the guidance).

ISBN: 978-1-4731-7668-3